Extracting Human Protein Information from MEDLINE Using a Full-Sentence Parser

Róbert Busa-Fekete, Kornél Kovács, and András Kocsor

Nowadays the MEDLINE [1] database is becoming the most comprehensive biomedical abstract repository among the life sciences literature. Due to its easy access and availability, it is one of the most widely-used sources of scientific data that uses several information retrieval systems. The NLM (U.S. National Library of Medicine) maintained MEDLINE database contains over 13 million references from about 4900 journals dating from 1965 to the present, and it is updated weekly. Obviously it is a crucial task in bioinformatics text mining to develop an automatic system that extracts information about genes and their interactions. That is why we were motivated in building an information extraction (IE) system that makes use of natural language processing (NLP) techniques.

In biological studies, the researchers are mostly interested in the interactions of the genes, so in this area of science biologists require an IE system that can search for relationships among human proteins [2]. That is why we will focus here only on genes that occur in living human cells. The National Center for Biotechnology Information (NCBI) [3] has many databases about gene interactions distributed taxonomically. Using these data sets we can easily obtain a subset of MEDLINE containing information about human gene interactions. We assume that in each observed abstract the gene names occur in it. We used a thesaurus containing about 58,000 gene names and their synonyms in order to annotate the gene names in the abstracts. The thesaurus was built up using three sources: Unified Medical Language System (UMLS) Metathesaurus [4], UMLS SPECIALIST Lexicon[4] and the Agilent Technologies[5] database.

With our approach we would like to learn more about the interactions of genes using full-sentence parsing [6, 7, 8]. Given a sentence, the syntactic parser assigns to it a syntactic structure, which consists of a set of labelled links connecting pairs of words. The parser also produces a constituent representation of a sentence (showing noun phrases, verb phrases, and so on). Using the syntactic information of each abstract, the biological interactions of genes can be predicted. Our IE system can handle certain types of gene interactions with the help of machine learning (ML) [9] methodologies (Artificial Neural Network [10], Support Vector Machines [11, 12]). Actually, many features of a syntactic tree can be represented as a multidimensional vector (i.e. depth and frequencies of different labels).

The performance of the IE process is influenced mostly by the quality of the syntactic parsing, which is why we chose to examine several methods to see how well they perform. A traditional approach, namely the Link Parser [13], will be applied as the baseline system. Because the LINK Parser is a general purpose syntax analyzer, its special biomedical extension will be also investigated. In addition, we propose a novel ML-based syntax parser for English. The algorithm interprets the words in a sentence as individual subtrees, and it concatenates the most suitable adjoining subtrees according to the ML model used.

When designing our system we had to take into account the fact that MEDLINE is a rapidly growing system and that the data is stored in compressed XML file format. So we created a framework which can handle the abstracts and their updates in their raw form, and incorporate them into our IE system. We evaluated our system on the well-known annotated Human Protein Reference Database (HPRD) [14] corpus and obtained some useful results.

References

[1] `http://www.pubmedcentral.nih.gov/`

[2] T. Sekimizu, H.S. Park and Jun'ichi. Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts, *Genome Informatics*, 9:62-71, 1998.

[3] http://www.ncbi.nlm.nih.gov/

[4] http://www.nlm.nih.gov/research/umls/

[5] http://www.home.agilent.com/

[6] D. Sleator and D. Temperley. Parsing English with a Link Grammar, *Carnegie Mellon University Computer Science technical report* CMU-CS-91-196, October 1991.

[7] J. Lafferty, D. Sleator, and D. Temperley. Grammatical Trigrams: A Probabilistic Model of Link Grammar, *Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language*, October, 1992.

[8] D. Grinberg, J. Lafferty, and D. Sleator. A robust parsing algorithm for link grammars, *Carnegie Mellon University Computer Science technical report* CMU-CS-95-125, and *Proceedings of the Fourth International Workshop on Parsing Technologies*, Prague, September, 1995.

[9] V. N. Vapnik. Statistical Learning Theory, *John Wiley and Son*, 1998.

[10] C.M. Bishop. Neural Networks for Pattern Recognition, *Oxford University Press*, 1995.

[11] N. Cristianini and J. Shawe-Taylor. Support Vector Machines and other kernel-based learning methods, *Cambridge University Press*, ISBN 0-521-78019-5, 2000.

[12] B. Schölkopf, C.J.C. Burges, and A.J. Smola. Advances in Kernel Methods: Support Vector Learning, *MIT PRess*, Cambridge, MA, 1999.

[13] http://www.link.cs.cmu.edu/link/

[14] http://www.hprd.org/